

eCuration: speed curating with PubTator

Zhiyong Lu, Earl Stadtman Investigator

National Center for Biotechnology Information (NCBI)
National Library of Medicine (NLM)
National Institutes of Health (NIH)



eCuration (computer-assisted biocuration) is necessary

Nature **455**, 47-50 (4 September 2008) | doi:10.1038/455047a; Published online 3 September 2008

Big data: The future of biocuration

Doug Howe¹, Maria Costanzo², Petra Fey³, Takashi Gojobori⁴, Linda Hannick⁵, Winston Hide^{6,7}, David P. Hill⁸, Renate Kania⁹, Mary Schaeffer^{10,11}, Susan St Pierre¹², Simon Twigger¹³, Owen White¹⁴ & Seung Yon Rhee¹⁵

To thrive, the field that links biologists and their data urgently needs structure, recognition and support.

▲ Top

The exponential growth in the amount of biological data means that revolutionary measures are needed for data management, analysis and accessibility. Online databases have become important avenues for publishing biological data. Biocuration, the activity of organizing, representing and making biological information accessible to both humans and computers, has become an essential part of biological discovery and biomedical research. But curation increasingly lags behind data generation in funding, development and recognition.



We propose three urgent actions to advance this key field. First, authors, journals and curators should immediately begin to work together to facilitate the exchange of data between journal publications and databases. Second, in the next five years, curators, researchers and university administrations should develop an accepted recognition structure to facilitate community-based curation efforts. Third, curators, researchers, academic institutions and funding agencies should, in the next ten years, increase the visibility and support of scientific curation as a professional career.

Bioinformatics (Oxford, England)

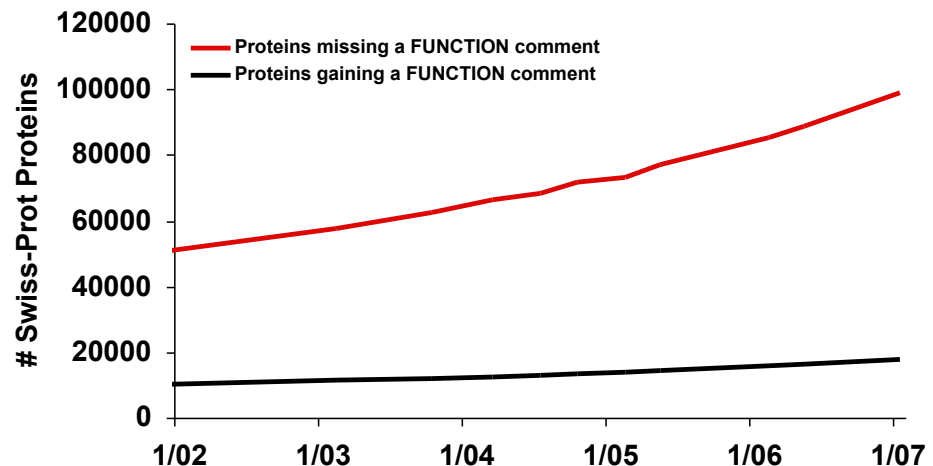
Author Manuscript

NIH Public Access

Manual curation is not sufficient for annotation of genomic databases

William A. Baumgartner, Jr, K. Bretonnel Cohen, [...], and Lawrence Hunter

[Additional article information](#)



Original article

Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II

Zhiyong Lu¹ and Lynette Hirschman^{2,*}

¹National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894 and ²The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA

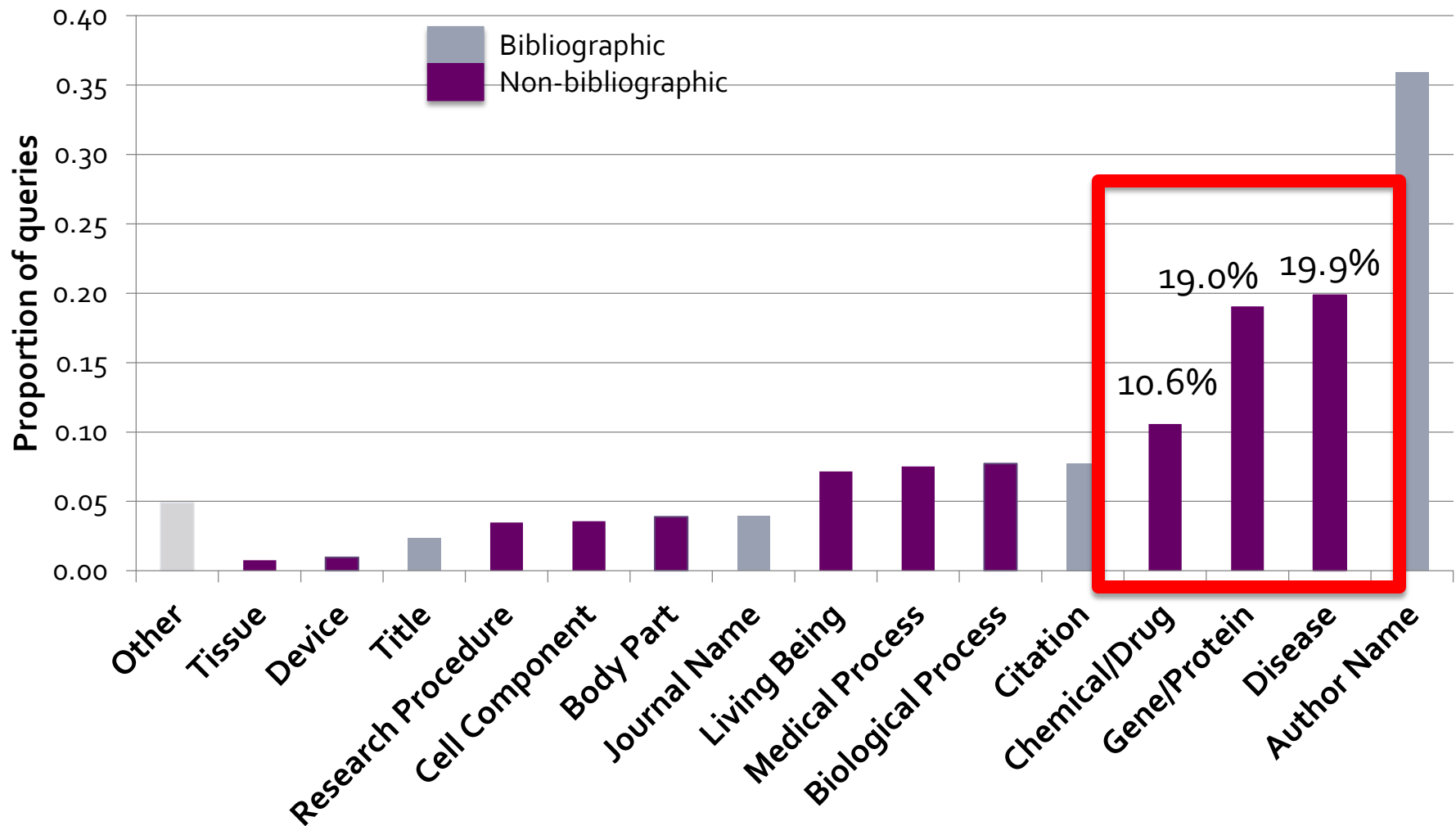
*Corresponding author: Tel: +1 781 271 7789; Fax: +1 781 271 2780; Email: lynette@mitre.org

Submitted 18 June 2012; Revised 24 August 2012; Accepted 2 October 2012

Manual curation of data from the biomedical literature is a rate-limiting factor for many expert curated databases. Despite the continuing advances in biomedical text mining and the pressing needs of biocurators for better tools, few existing text-mining tools have been successfully integrated into production literature curation systems such as those used by the expert curated databases. To close this gap and better understand all aspects of literature curation, we invited submissions of written descriptions of curation workflows from expert curated databases for the BioCreative 2012 Workshop Track II. We received seven qualified contributions, primarily from model organism databases. Based on these descriptions, we identified commonalities and differences across the workflows, the common ontologies and controlled vocabularies used and the current and desired uses of text mining for biocuration. Compared to a survey done in 2009, our 2012 results show that many more databases are now using text mining in parts of their curation workflows. In addition, the workshop participants identified text-mining aids for finding gene names and symbols (gene indexing), prioritization of documents for curation (document triage) and ontology concept assignment as those most desired by the biocurators.

Database URL: <http://www.biocreative.org/tasks/bc-workshop-2012/workflow/>

Most searched topics in PubMed



Neveol, Dogan, Lu, Semi-automatic semantic annotation of PubMed queries:
A study on quality, efficiency, satisfaction, *Journal of Biomedical Informatics*, 2010

Key biological entities

Disease

- diabetes mellitus; DM; type 2 diabetes

Genomic variation

- c.77A>C; c.77A->C; A77C; AC

Gene/Protein

- TP53; tumor protein p53; p53; BCC7; LFS1

Species

- Arabidopsis thaliana; thale-cress; AT

Chemical/Drug

- Aspirin; 2-(Acetyloxy)benzoic Acid; Acetysal

Our NER Tools



Disease

DNorm – 80.90%



Mutation

tmVar – 91.39%

Gene/Protein

GenNorm – 84.50%



Species

SR4GN – 85.42%

Chemical/Drug

tmChem – 88.27%



- Freely available & open source
- High Performance
 - DNorm: Best in 2013 ShARe/CLEF shared task on Disease Normalization
 - tmChem: Best in 2013 BioCreative IV Chemical Entity Mention task
 - GenNorm: Best in 2010 BioCreative III Gene Normalization Task
- BioC format compatible for improved interoperability

All numbers are F1 scores

Our tmTools are publicly available

To make it easy for biocurators, we have already applied all these tools to PubMed abstracts and store results in our Web-based annotation tool – PubTator!

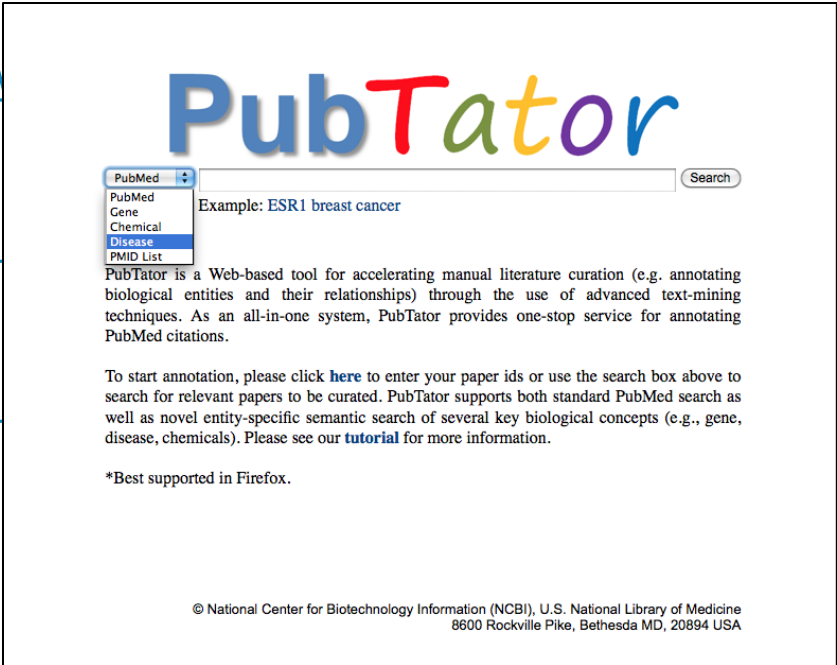
■ **SR4GN:** [www](http://www.ncbi.nlm.nih.gov/downloads/SR4GN/)

■ **GenNorm:** h

■ **tmChem:** [htt](http://h/Lu/Demo/tmChem/)

[wnloads/SR4GN/](http://www.ncbi.nlm.nih.gov/downloads/SR4GN/)

h/Lu/Demo/tmChem/



PubTator

PubMed
Gene
Chemical
Disease
PMID List

Example: ESR1 breast cancer

Search

PubTator is a Web-based tool for accelerating manual literature curation (e.g. annotating biological entities and their relationships) through the use of advanced text-mining techniques. As an all-in-one system, PubTator provides one-stop service for annotating PubMed citations.

To start annotation, please click [here](#) to enter your paper ids or use the search box above to search for relevant papers to be curated. PubTator supports both standard PubMed search as well as novel entity-specific semantic search of several key biological concepts (e.g., gene, disease, chemicals). Please see our [tutorial](#) for more information.

*Best supported in Firefox.

© National Center for Biotechnology Information (NCBI), U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

PubTator Intro/Highlights

1. Web-based; no installation required; in sync with PubMed
2. One-stop curation service from literature search to annotation
3. Curator friendly (PubMed-like) interface; easy to use
4. Integrates competition-winning text-mining tools for automatic pre-annotations
5. Easy to adapt and customize to different curation tasks



Wei, Kao, & Lu: PubTator: a Web-based text-mining tool for assisting biocuration, to appear in Nucleic Acids Research, 2013

PubTator's Curation Interface

Go back

☐ Curatable
☐ Not Curatable

Document triage

Bioconcepts

☒ Disease ☒ Species ☒ Chemical ☒ Gene

PMID:23546941

Decreased expression of the **DBC2** gene and its clinicopathological significance in **breast cancer**: correlation with aberrant DNA methylation.

Publication: Biotechnology letters; 2013 Apr 2

Gene Chemical Disease Species Clear Reset

TITLE:

Decreased expression of the **DBC2** gene and its clinicopathological significance in **breast cancer**: correlation with aberrant DNA methylation.

ABSTRACT:

Loss of **DBC2** (deleted in **breast cancer** 2) gene expression is frequent in **breast cancer** tissues. This can be explained by homozygous deletions or other mutations in a minority of cases but alternative mechanisms need to be investigated. Here, **DBC2** expression was significantly suppressed compared with normal breast tissues in **breast cancer** tissues when analyzed by RT-PCR. Furthermore, DNA methylation on **DBC2** was more prevalent in **breast tumors** than in normal tissues. **DBC2** mRNA levels correlated with the degree of **DBC2** methylation in **breast cancer** tissues and in a **breast cancer** cell line (T47D). Clinico-pathological correlation analysis showed that **DBC2** promoter methylation was associated with tumor-node-metastasis stages II and III/IV, lymph node metastasis, p53 mutation, and **HER2**-positive status. Thus loss of **DBC2** expression is caused by abnormal methylation of **DBC2** and might have a role in **breast cancer** development.

☒ Concept View ☐ Mention View [Add bio-relation annotation to the table below.](#)

Entity type	Entity mention	Concept ID	Nomenclature	GD	Delete
Disease	breast cancer breast tumors	D001943	MEDIC	<input type="checkbox"/>	Delete
Gene	DBC2	23221	NCBI Gene	<input type="checkbox"/>	Delete
Gene	HER2	2064	NCBI Gene	<input type="checkbox"/>	Delete
Disease	metastasis	D009362	MEDIC	<input type="checkbox"/>	Delete
Disease	tumor	D009369	MEDIC	<input type="checkbox"/>	Delete

Bio-concept annotation

Relation name	Relation type	Bio-entities	Delete
GD	Gene_Disease	DBC2 breast cancer	Delete

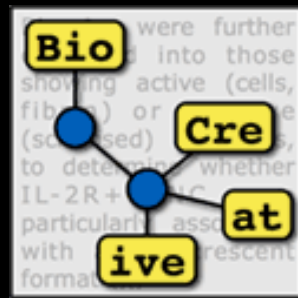
Bio-relation annotation

Save Annotation Results

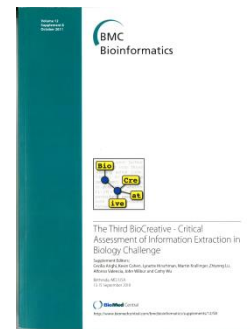
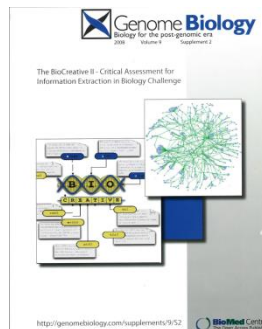
Save & Export Annotation Results

BioCreative Challenge (2003 –)

www.biocreative.org



BioCreative	Workshop Location	Workshop date	GM	GN	GO	PPI	IAT	CTD	Curation Workflow	BioC	CHEM DNER
BC I	Granada, Spain	Mar, 2004	●	●	●						
BC II	Madrid, Spain	Apr, 2007	●	●		●					
BC II.5	Madrid, Spain	Oct, 2009				●					
BC III	Bethesda, USA	Sep, 2010		●		●	●				
BC 2012	DC, USA	Oct, 2012					●	●	●		
BC IV	Bethesda, USA	Oct, 2013			●		●	●		●	●



PubTator evaluation



- Task: manually annotating genes in 50 abstracts
- Experimental settings (25 abstracts each)
 1. PubMed + spreadsheet (baseline)
 2. PubTator + computer-generated gene results
- Results: 40% decrease in curation time & slightly higher accuracy

Wei, Harris, ... Lu. Accelerating literature curation with text mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database*, 2012; bas041

Top rated by biocurators

Table 8

Overall rating for each system by category


Subjective measure (overall median for each section)

System	Overall evaluation	Task completion	System design	Learnability	Usability	Recommendation
PubTator	6.0	5.5	6.0	6.0	6.5	7.0
eFIP	6.0	6.0	6.0	6.0	7.0	5.5
Tagtog ^a	5.0	5.0	5.0	5.0	6.0	4.5
Textpresso	4.0	5.0	5.0	5.0	6.0	4.5
PCS	4.0	3.0	6.0	6.0	6.0	4.0
PPInterFinder	4.0	2.5	5.0	5.0	5.0	3.0
T-HOD	4.0	3.0	4.0	5.0	5.0	3.0

Median for questions linked for each of the categories. Likert scale from 1 to 7, from worst to best, system was only reviewed at the workshop.



Arighi, et al., An overview of the BioCreative 2012 Workshop Track III: interactive text mining task. *Database*, 2013. bas056

Successful applications of PubTator

















HuGE Navigator (version 2.0)
 An integrated, searchable knowledge base of genetic associations and human genome epidemiology.


[Home](#) | [Download Center](#) | [Open Source Projects](#) | [Contact](#)

Curator pick of the week:
 Surgical management of breast cancer in BRCA-mutation carriers: a systematic review and meta-analysis. Valachis A, Nearchou AD, Lind P. Breast Cancer Res Treat. 2014 Feb 25
[PubMed Link](#)

HuGE Navigator is a continuously updated knowledge base in human genome epidemiology, including population prevalence of genetic variants, genetic associations... [more](#)
 Join us on Facebook  and follow us on Twitter 
 Last database update: **07 Mar 2014**


Site citation: W Yu, M Gwinn, M Clyne, A Yesupriya & M J Khoury. *A Navigator for Human Genome Epidemiology. Nat Genet* 2008 Feb;40(2): 124-5.

 Phenopedia: Look up genetic associations and human genome epidemiology summaries by disease.	 Genopedia: Look up genetic associations and human genome epidemiology summaries by gene.
 HuGE Literature Finder: Find published articles in genetic associations and human genome epidemiology.	 Gene Prospector: A gateway for evaluating genes in relation to disease and risk factors.
 GWAS Integrator: Explore published GWAS and relevant information.	 Cancer GAMAdb: Database of cancer genetic associations from meta analyses and GWAS.
 HuGE Watch: Track the evolution of published literature in human genome epidemiology.	 Variant Name Mapper: Map common names and rs numbers of genetic variants.
 HuGE Investigator Browser: Find investigators in a particular field of human genome epidemiology.	 Genotype Prevalence Catalog: Present genotype estimates in US population.
 Download Center: Download complete datasets from different databases/applications.	 GAPscreener: Screening tool for published literature genetic associations.
 HuGE Risk Translator: Calculate the predictive value of genetic markers for disease risk.	 Open Source: Infrastructure for managing knowledge information from PubMed.
 HuGE Track : A custom track built for HuGE data in the UCSC Genome Browser.	


 © 2010 HuGE Navigator All rights reserved.

[Home](#) | [HuGENet™](#) | [Open Source Projects](#) | [Site Map](#) | [Contact](#)

“PubTator substantially reduces the manual data input involved, reflected in both time-savings and reduction in physical fatigue of keyboard typing.” – Mindy C.



[Journal home](#) > [Archive](#) > [Correspondence](#) > [Full Text](#)

Journal content

- Journal home
- Advance online publication
- Current issue
- Archive**
- Focuses and Supplements
- Press releases

Correspondence
Nature Genetics **40**, 124 - 125 (2008)
 doi:10.1038/ng0208-124
A navigator for human genome epidemiology
 Wei Yu¹, Marta Gwinn¹, Melinda Clyne¹, Ajay Yesupriya¹ & Muin J Khoury¹
 1. National Office of Public Health Genomics Centers for Disease Control and Prevention, Atlanta, Georgia 30309, USA.
 Correspondence to: Wei Yu¹ e-mail: wby0@cdc.gov.

Discussions

- eCuration: computer-assisted curation can improve productivity
- Future directions
 - Working with ontologies
 - Working with full-text
- What would you do with PubTator?

Acknowledgments

■ My Team

- Rezarta Dogan
- Bethany Harris
- Ritu Khare
- Aurelie Neveol
- Yuqing Mao
- Robert Leaman
- Jiao Li
- Chih-Hsuan Wei

■ BioCreative

- Lynette Hirschman, MITRE
- Kevin Cohen, U of Colorado
- Alfonso Valencia; Martin Krallinger, CNIO
- Cecilia Arighi, Cathy Wu, U of Delaware
- Carolyn Mattingly; Tom Wiegers, NCSU



Pacific Symposium on Biocomputing (PSB) 2015

January 4 – 8, 2015

The Big Island of Hawaii

Crowdsourcing and Mining Crowd Data

Crowdsourcing techniques
microtask environments
games with a purpose
workflow sequestration

Crowd data
human genomics sequence data
electronic health records
social media data

Robert Leaman and Zhiyong Lu, NCBI/NLM/NIH
Ben Good and Andrew Su, Scripps Research Institute

Questions?

Thank you!

zhiyong.lu@nih.gov